CS 6824: Application of Basic Clustering Algorithms to Find Expression Modules in Cancer

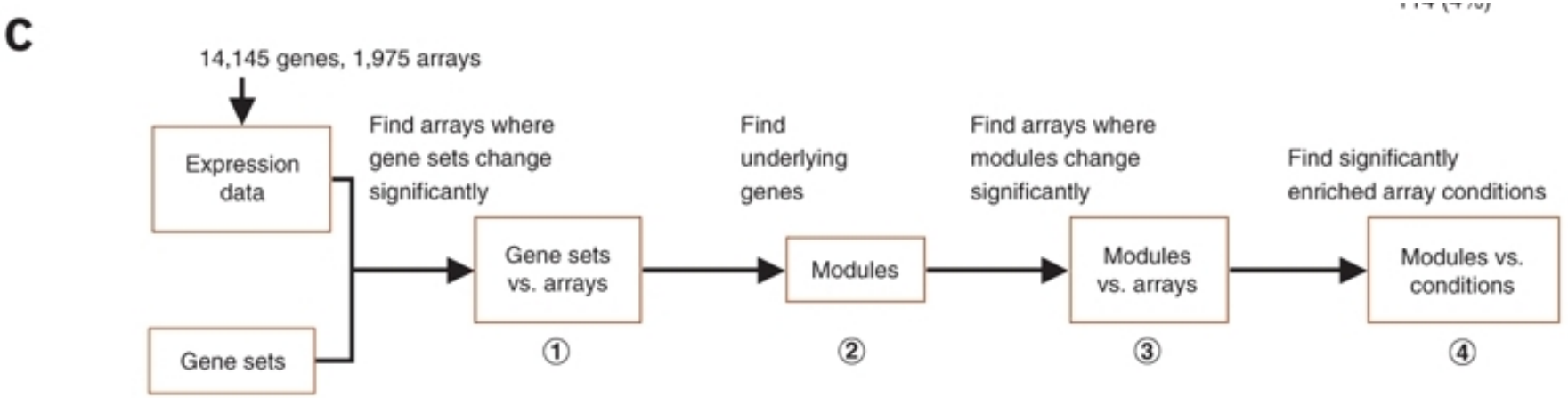
T. M. Murali

February 16, 2011

Innovative Application of Hierarchical Clustering

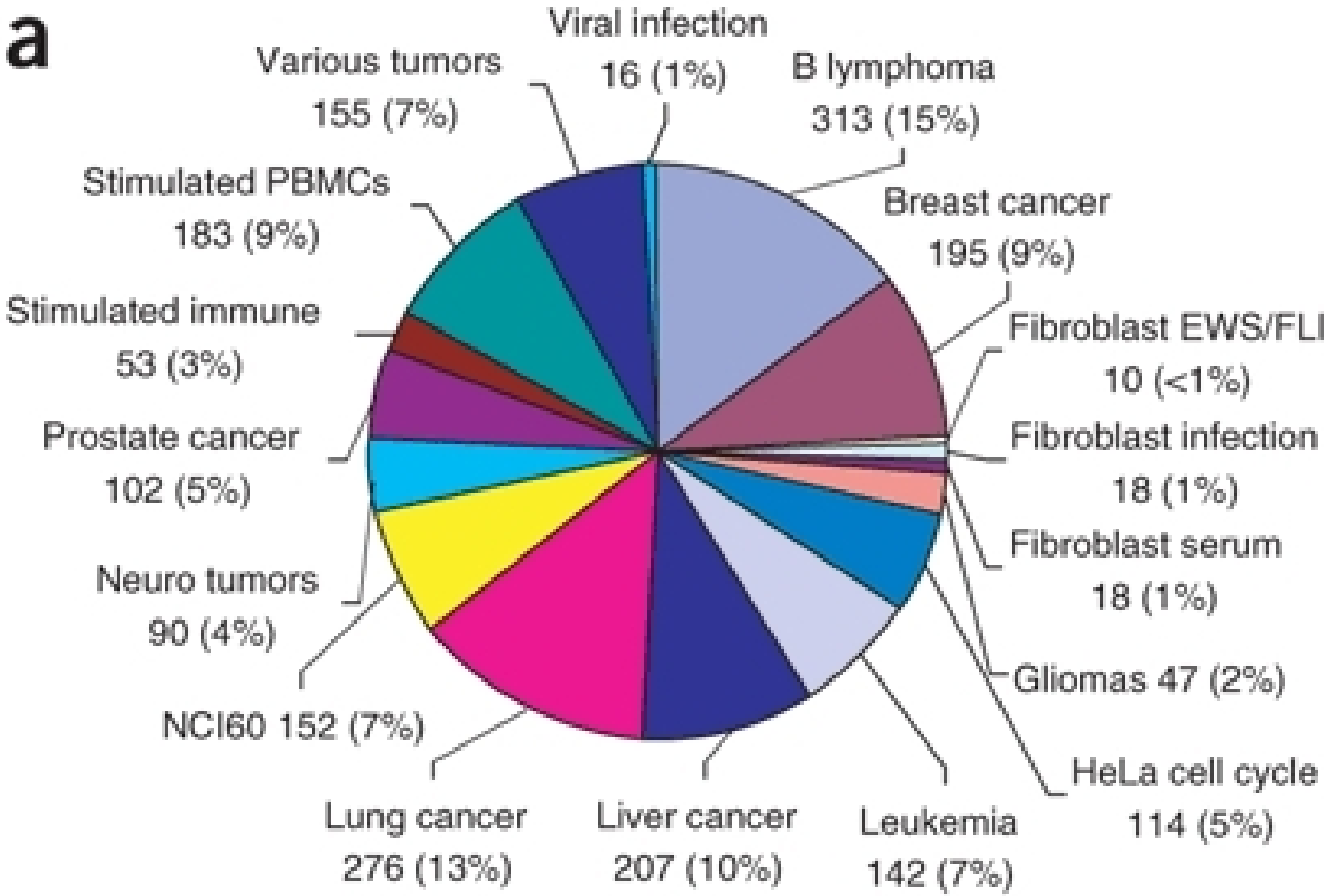
- *A module map showing conditional activity of expression modules in cancer*, Eran Segal, Nir Friedman, Daphne Koller and Aviv Regev, Nature Genetics 36, 1090–1098, 2004
- Analyse gene expression data to find groups of genes expressed in concert between different cancers.
- Use hierarchical clustering innovatively.

Key Ideas



- Group genes into predefined *gene sets*, e.g., groups of genes with the same functional annotation.
- Convert gene-by-array matrix into gene-set-by-array matrix.
- Hierarchically cluster gene sets in this matrix.
- Identify "interesting" gene set clusters (nodes) in the tree.
- In each gene set cluster, remove genes not expressed consistently with the cluster.

Gene Expression Data Sets



Data Normalisation

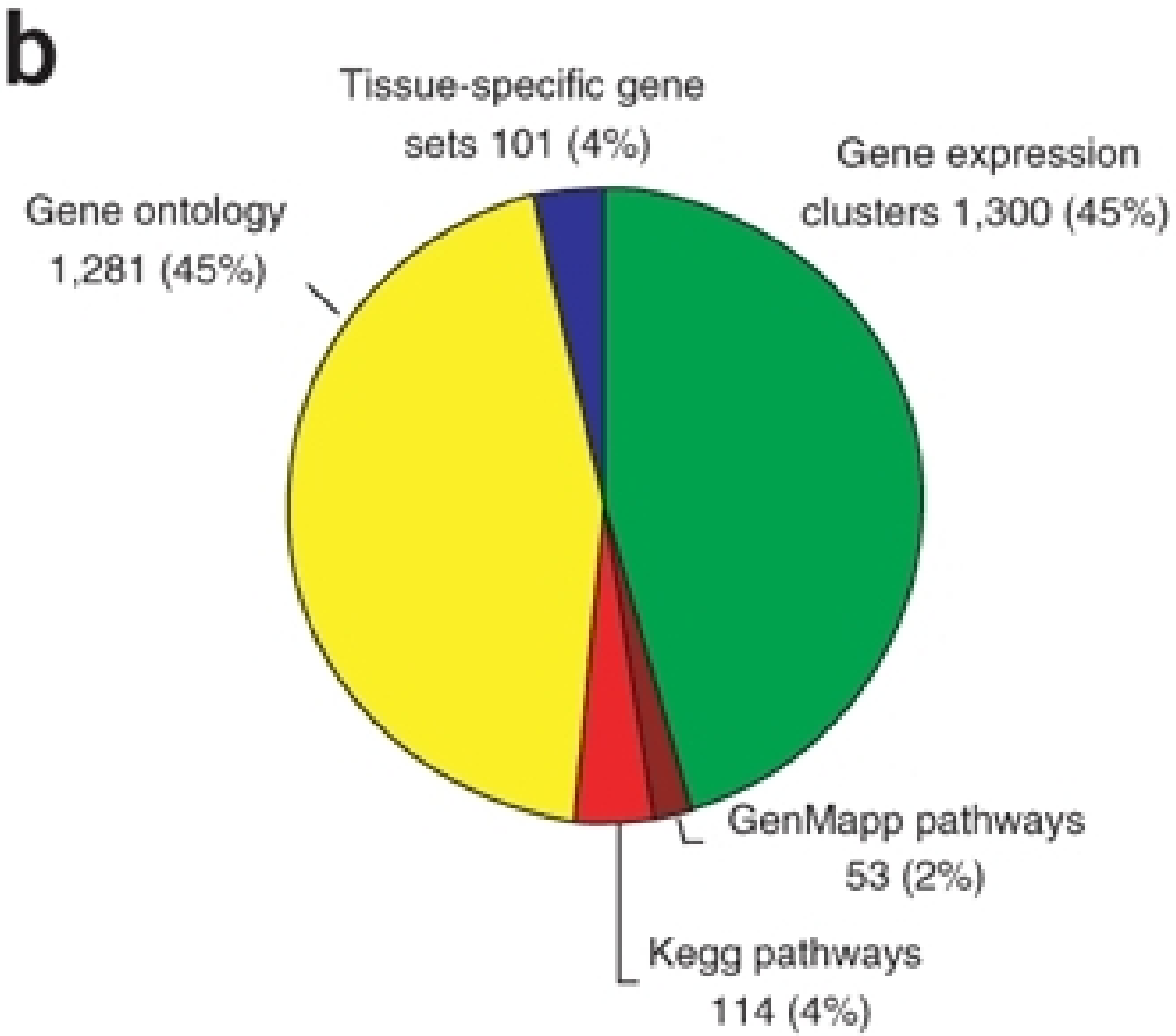
Data Normalisation

- Needed because some arrays measure "absolute" value of gene expression and others measure "relative" values.
- Affymetrix microarrays: take logarithm to the base-2 and zero transform within data set.
- cDNA microarrays:

Data Normalisation

- Needed because some arrays measure "absolute" value of gene expression and others measure "relative" values.
- Affymetrix microarrays: take logarithm to the base-2 and zero transform within data set.
- cDNA microarrays: zero transform within data set.

Pre-defined Gene Sets



Computing Gene-Set-By-Array Matrix

- Goal is to construct a gene-set-by-array matrix.
- For each gene set-array pair, find an "average" expression value of that gene set in that array.

Computing Gene-Set-By-Array Matrix

- Goal is to construct a gene-set-by-array matrix.
- For each gene set-array pair, find an "average" expression value of that gene set in that array.
- A gene is *induced* (respectively, *repressed*) in an array if its log-fold change in expression is ≥ 1 (respectively, ≤ -1).
- For each gene set-array pair, compute the fraction of genes induced or repressed.
- Use these values in the gene-set-by-array matrix.

Computing Significant Entries in the Gene-Set-By-Array Matrix

- Many entries in the gene-set-by-array matrix may not be statistically significant.

Computing Significant Entries in the Gene-Set-By-Array Matrix

- Many entries in the gene-set-by-array matrix may not be statistically significant.
- For a given array, fraction of induced genes in a gene set may be close to the fraction of induced genes in the array.

Computing Significant Entries in the Gene-Set-By-Array Matrix

- Many entries in the gene-set-by-array matrix may not be statistically significant.
- For a given array, fraction of induced genes in a gene set may be close to the fraction of induced genes in the array.
- Statistical test: for a given array, is the fraction of induced genes in a gene set much larger than the fraction of induced genes in the entire array?

Computing Significant Entries in the Gene-Set-By-Array Matrix

- Many entries in the gene-set-by-array matrix may not be statistically significant.
- For a given array, fraction of induced genes in a gene set may be close to the fraction of induced genes in the array.
- Statistical test: for a given array, is the fraction of induced genes in a gene set much larger than the fraction of induced genes in the entire array?
- Compute the p-value of the fraction using Fisher's exact test.
- Do so for every gene-set-array pair.
- Use false discovery rate correction to account for testing multiple hypotheses.
- Replace insignificant entries by 0.

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.
- Informally, $u_{G,a}/m_G \approx u_a/m$ is not statistically significant.

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.
- Informally, $u_{G,a}/m_G \approx u_a/m$ is not statistically significant.
- Formally, what is the probability that if we pick m_G genes at random from m genes, we will select $u_{G,a}$ or more that are induced in a?

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.
- Informally, $u_{G,a}/m_G \approx u_a/m$ is not statistically significant.
- Formally, what is the probability that if we pick m_G genes at random from m genes, we will select $u_{G,a}$ or more that are induced in a?

$$\sum_{i \geq u_{G,a}}$$

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.
- Informally, $u_{G,a}/m_G \approx u_a/m$ is not statistically significant.
- Formally, what is the probability that if we pick m_G genes at random from m genes, we will select $u_{G,a}$ or more that are induced in a?

$$\sum_{i \geq u_{G,a}} \frac{\binom{u_a}{i}\binom{m-u_a}{m_G-i}}{\binom{m}{m_G}}$$

Computing the Significance of an Entry in the Gene-Set-By-Array Matrix

- Let m be the number of genes in the data set.
- let m_G be the number of genes in a gene set G.
- Let u_a be the number of induced genes in an array a.
- let $u_{G,a}$ be the number of genes in G induced in a.
- Informally, $u_{G,a}/m_G \approx u_a/m$ is not statistically significant.
- Formally, what is the probability that if we pick m_G genes at random from m genes, we will select $u_{G,a}$ or more that are induced in a?

$$\sum_{i \geq u_{G,a}} \frac{\binom{u_a}{i}\binom{m-u_a}{m_G-i}}{\binom{m}{m_G}}$$

- If this probability is at most a user-specified threshold, we deem that entry to be statistically significant.

Hierarchical Clustering

- Start from a gene-set-by-array matrix containing fraction of induced/repressed genes. Fraction is negative if repressed.
- Apply bottom-up hierarchical clustering.
- Vector at internal node is

Hierarchical Clustering

- Start from a gene-set-by-array matrix containing fraction of induced/repressed genes. Fraction is negative if repressed.
- Apply bottom-up hierarchical clustering.
- Vector at internal node is average of vectors at descendant leaves.

Hierarchical Clustering

- Start from a gene-set-by-array matrix containing fraction of induced/repressed genes. Fraction is negative if repressed.
- Apply bottom-up hierarchical clustering.
- Vector at internal node is average of vectors at descendant leaves.
- Which nodes do we select as clusters in the tree?

Hierarchical Clustering

- Start from a gene-set-by-array matrix containing fraction of induced/repressed genes. Fraction is negative if repressed.
- Apply bottom-up hierarchical clustering.
- Vector at internal node is average of vectors at descendant leaves.
- Which nodes do we select as clusters in the tree?
  - Associate each interior node with Pearson correlation between the two children.
  - Cluster $\equiv$ node whose Pearson correlation differs by more than 0.05 from the Pearson correlation of its parent.

Turning Clusters into Modules

- Each cluster is the union of descendant gene sets (leaves).
- Module $\equiv$ Cluster minus genes whose expression is not consistent with the rest of the cluster.

Testing Consistency of a Gene with a Gene Set

- Let g be the gene and G be the gene set.

Testing Consistency of a Gene with a Gene Set

- Let g be the gene and G be the gene set.
- Let I (respectively, R) be the set of arrays in which G is significantly induced (respectively, repressed).
- For an array a in I (or R), let p_a be the fraction of genes that are induced (or repressed) by two-fold or more in a.

Testing Consistency of a Gene with a Gene Set

- Let g be the gene and G be the gene set.
- Let I (respectively, R) be the set of arrays in which G is significantly induced (respectively, repressed).
- For an array a in I (or R), let p_a be the fraction of genes that are induced (or repressed) by two-fold or more in a.
- Measure extent to which g's expression changed by more (or less) than two-fold in the arrays in I (or R):

Testing Consistency of a Gene with a Gene Set

- Let g be the gene and G be the gene set.
- Let I (respectively, R) be the set of arrays in which G is significantly induced (respectively, repressed).
- For an array a in I (or R), let p_a be the fraction of genes that are induced (or repressed) by two-fold or more in a.
- Measure extent to which g's expression changed by more (or less) than two-fold in the arrays in I (or R):

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

Testing Consistency of a Gene with a Gene Set

- Let g be the gene and G be the gene set.
- Let I (respectively, R) be the set of arrays in which G is significantly induced (respectively, repressed).
- For an array a in I (or R), let p_a be the fraction of genes that are induced (or repressed) by two-fold or more in a.
- Measure extent to which g's expression changed by more (or less) than two-fold in the arrays in I (or R):

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

- No contribution from an array in I (or R) is g is not induced (or not repressed) in a.
- Larger contribution from arrays with fewer induced genes.
- Compute statistical significance of this score.

Computing Statistical Significance of Score(g)

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

Computing Statistical Significance of Score(g)

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

- Null hypothesis: genes in each array are randomly permuted, i.e., the p_a induced genes in an array $a \in I$ are chosen randomly.

Computing Statistical Significance of Score(g)

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

- Null hypothesis: genes in each array are randomly permuted, i.e., the p_a induced genes in an array $a \in I$ are chosen randomly.
- Each element in Score(g) is an independent binary random variable.
- Random variable takes the value $-\log(p_a)$ with probability p_a and the value 0 with the probability $1 - p_a$.

Computing Statistical Significance of Score(g)

$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

- Null hypothesis: genes in each array are randomly permuted, i.e., the p_a induced genes in an array $a \in I$ are chosen randomly.
- Each element in Score(g) is an independent binary random variable.
- Random variable takes the value $-\log(p_a)$ with probability p_a and the value 0 with the probability $1 - p_a$.
- Mean of Score(g) is $\sum_{a \in I \cup R} -p_a \log p_a$ and variance is $\sum_{a \in I \cup R} p_a(1 - p_a) \log^2 p_a$.

Computing Statistical Significance of Score(g)

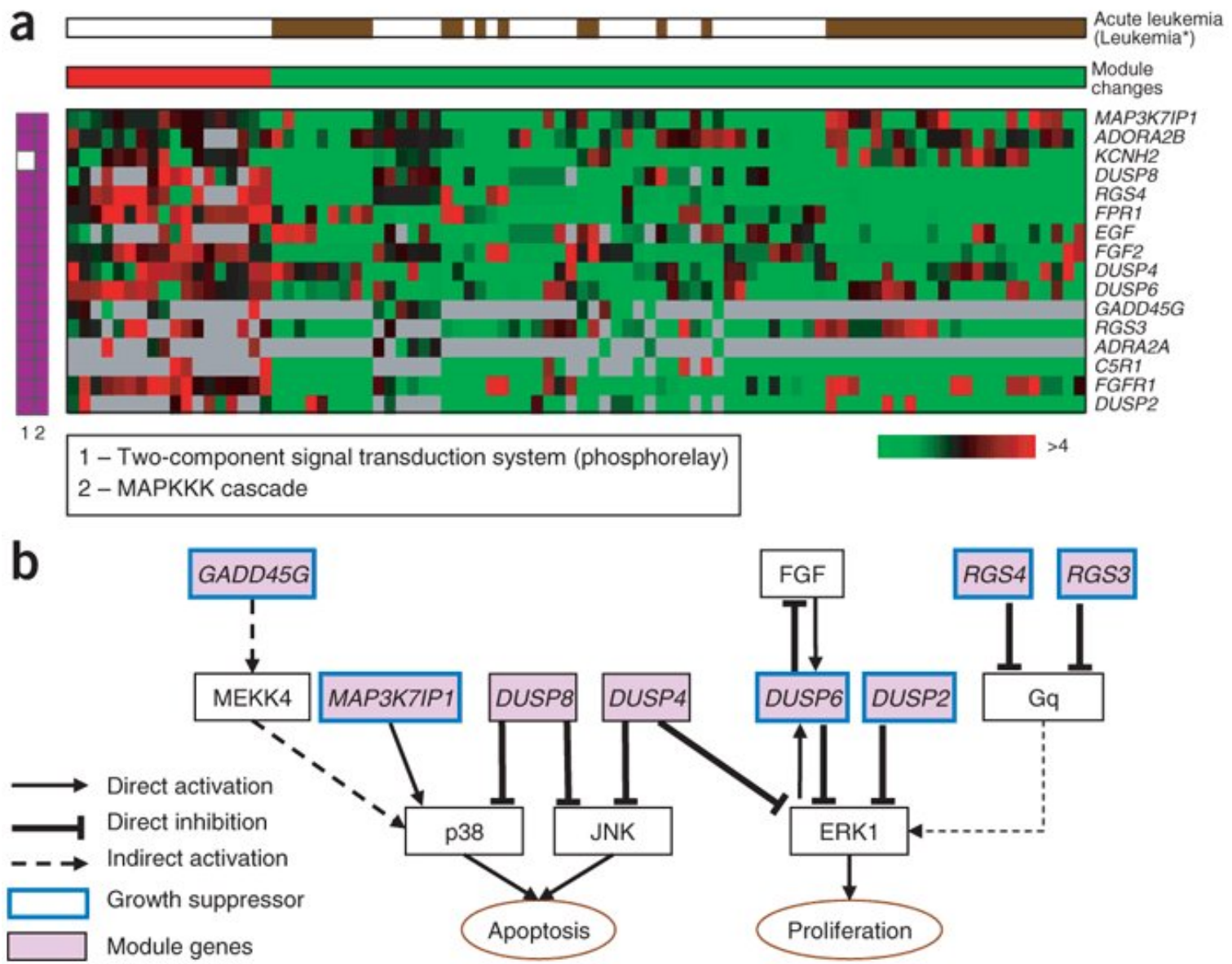
$$\text{Score}(g) = \sum_{a \in I | g \text{ is induced in } a} -\log(p_a) + \sum_{a \in R | g \text{ is repressed in } a} -\log(p_a)$$

- Null hypothesis: genes in each array are randomly permuted, i.e., the p_a induced genes in an array $a \in I$ are chosen randomly.
- Each element in Score(g) is an independent binary random variable.
- Random variable takes the value $-\log(p_a)$ with probability p_a and the value 0 with the probability $1 - p_a$.
- Mean of Score(g) is $\sum_{a \in I \cup R} -p_a \log p_a$ and variance is $\sum_{a \in I \cup R} p_a(1 - p_a) \log^2 p_a$.
- Central limit theorem $\Rightarrow$ that the distribution of Score(g) is well-approximated by a Gaussian distribution with this mean and variance.
- Assess statistical significance by computing the tail of this Gaussian.

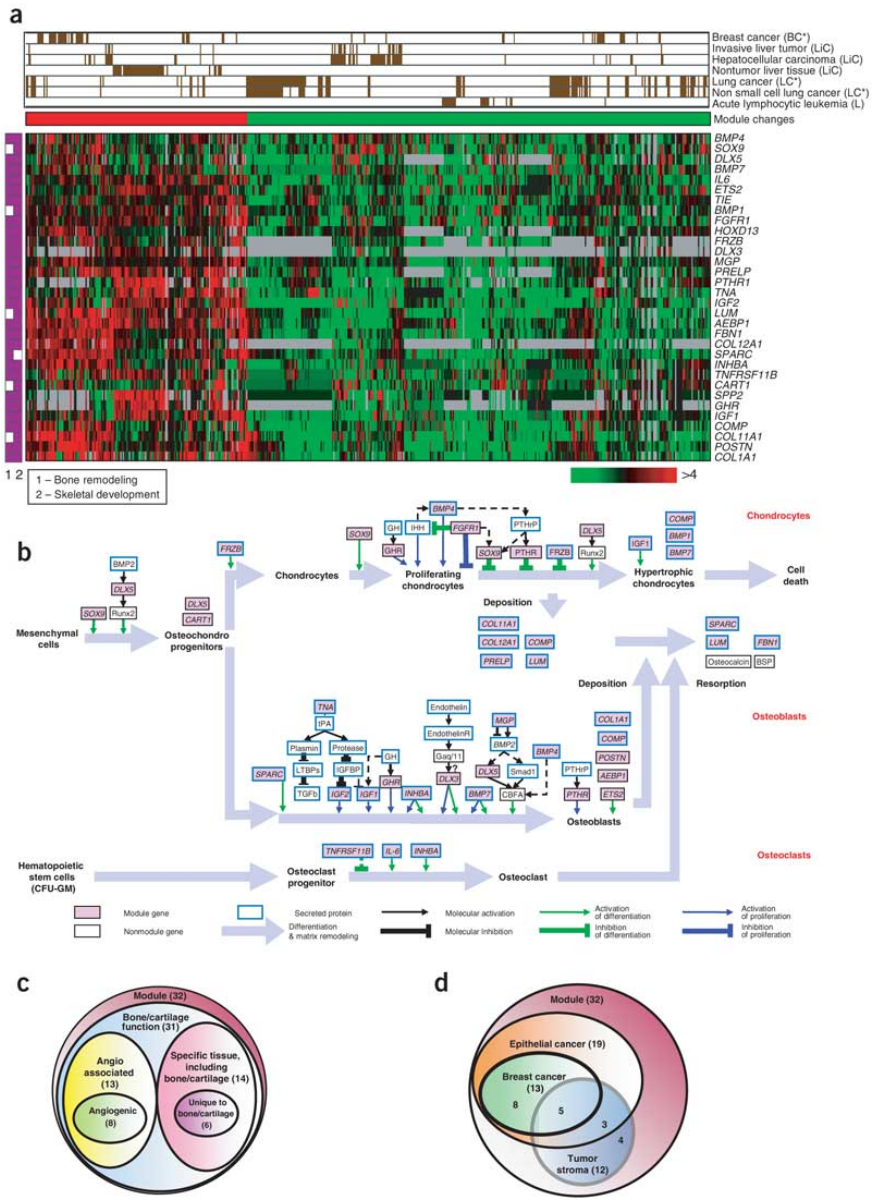
Further Analysis

- Statistical significance of computed modules using leave-one-out cross validation (read supplement).
- Compute enrichment of clinical annotations of the arrays in a module.
- Visualisation of modules.
- Literature-based analysis of modules

Growth Inhibitory Module



Bone Osteoblastic Module



Conclusions

- Used pre-defined gene sets to drive hierarchical clustering algorithm.
- Remove genes from a cluster of gene sets if the gene's expression profile deviates from the cluster.
- Automatically decide which arrays are part of a module.
- Interpretation of each module requires a lot of manual analysis.
- Such manual analysis is required for deciding subsequent experiments.